

RESEARCH ON LINEAR SYSTEMS OF A VERY LARGE SIZE

/ INTERIM REPORT

for the period

6 May to 15 January 1967

Prepared for

National Aeronautics and Space Administration
Washington, D. C.

Prepared by

/ Raytheon Company
Space and Information Systems Division
Autometric Operation
4217 Wheeler Avenue
Alexandria, Virginia

FACILITY FORM 602

N67-25495

(ACCESSION NUMBER)

56
(PAGES)

CR-83789
(NASA CR OR TMX OR AD NUMBER)

(THRU)

19
(CODE)

(CATEGORY)

TABLE OF CONTENTS

<u>Section</u>	<u>Title</u>	<u>Page</u>
1	INTRODUCTION.....	1
	1.1 Review and Evaluation of Major Procedures Now in Use.....	3
	1.2 Solution of Singular or Poorly Condition- ed Systems.....	3
	1.3 The Matrix as a Continuous Function.....	4
	1.4 The Matrix as a Geometric Object.....	5
	1.5 Specific Computational Difficulties.....	6
2	REVIEW AND EVALUATION OF PROCEDURE FOR SOLUTION OF LINEAR SYSTEMS.....	7
	2.1 Classification of Procedures.....	7
	2.2 Partition Procedures.....	16
	2.3 Enlargement Procedures.....	18
	2.4 Orthogonalization Procedures.....	21
	2.5 Indirect Procedures.....	23
	2.5.1 Universally Convergent Procedures.....	23
	2.5.2 Conditionally Convergent Procedures.....	25
3	THE HAWKINS UNIVERSALLY CONVERGENT INDIRECT PROCEDURE.....	31
4	THE MATRIX AS A CONTINUOUS PROCESS.....	34
5	INVESTIGATION OF TOPOLOGY OF LARGE MATRICES....	36
6	COMPUTATION PROCEDURES.....	44
	6.1 The Linear System.....	44
	6.2 Preliminary Handling and Computation.....	44
7	SUMMARY.....	46
APPENDIX I	BIBLIOGRAPHY	

LISTING OF TABLES

<u>Table</u>	<u>Title</u>	<u>Page</u>
1	Classification of Procedures.....	8
2	Direct Procedure.....	12
3	Matrix Inversion.....	14
4	Square Root Procedure.....	16
5	Conjugate Gradient Procedures.....	22
6	Universal Procedures.....	24
7	Iterative Procedures.....	27
8	C.C. Indirect Procedures.....	28
9	Computation of $A^T A$	42

1. INTRODUCTION

An investigation of the methods of solving large linear systems would not normally include an investigation into methods of finding the reciprocal of a matrix unless such reciprocals were directly applicable to the solution. Reciprocal matrices appear to have little use in solving large systems; and the number of steps required for a solution is greater and the avoidance of small divisors in the solving process is often difficult and sometimes impossible. Nevertheless, there is at least one good reason for including reciprocal matrices in the investigation; every problem of practical interest requires for a complete solution, not only the values of the unknowns, but also estimates of the reliabilities of these values, and the reliabilities are commonly expressed in terms of the elements of a reciprocal matrix. That is, if the problem requires that the vector X be found as a function of the observation vector Y , where:

$$Y = f(X) \quad , \quad (1)$$

a linear system of equations is derived (if f is not already linear)

by finding the matrix $A \equiv [\partial Y / \partial X]$ (2)

and solving, instead of (1), the simpler equation:

$$\Delta Y = A \Delta X \quad (3)$$

where:

$$\begin{aligned} Y &= Y \text{ (assumed)} + \Delta Y \\ X &= X \text{ (assumed)} + \Delta X \end{aligned} \quad (4)$$

If \bar{A}^{-1} is the generalized reciprocal of A (ref. 1,2,3)

$$\begin{aligned}\bar{A}^{-1} &\equiv (A^T A)^{-1} A^T \\ \bar{A} &\equiv A\end{aligned}, \quad (5)$$

then

$$\bar{A}^{-1} \Delta Y = \Delta X. \quad (6)$$

The variance Σ_X^2 of X is related to the variance Σ_Y^2 of Y by

$$\Sigma_X^2 = \bar{A}^{-1} \Sigma_Y^2 (\bar{A}^{-1})^T, \quad (7)$$

and this, for

$$\Sigma_Y^2 = I$$

gives

$$\Sigma_X^2 = (A^T A)^{-1}. \quad (8)$$

Since several solution procedures allow computation of $(A^T A)^{-1}$ almost simultaneously with computation of ΔX , and in such a way that some steps in the computation are the same for $(A^T A)^{-1}$ and ΔX , it seems natural to include an investigation of methods of computing \bar{A}^{-1} (or, more usually $[A^T A]^{-1}$) when such investigation forms a natural part of the main problem.

The theory of very large linear systems is a subset in the theory of linear systems. It has grown rapidly along with the growth in size and speed of computers able to use such theory, and now has an extensive literature of its own. (Appendix I gives a fairly complete bibliography of the literature through 1966; it is an expanded version of a bibliography submitted previously.) The subject is, in fact, so great that an investigation of limited range like the present cannot hope to even

start covering more than a small part of the subject. The investigation has, therefore, been restricted to those aspects of the subject which appeared most promising of new or useful results, or on which so little work had been done that further investigation was needed to find out what might result. In other words, to make efficient use of the limited effort allotted, this effort was channeled into the investigation of a small number of avenues which were relatively unexplained but held some promise of leading to useful results. The avenues chosen for investigation were the following.

1.1 Review and Evaluation of Major Procedures Now in Use

There is a vast variety of procedures given in the literature. They are most of them variants of a few major procedures, and differences between varieties within a major category are small when applied to a general problem although they may be large for particular varieties of problems. Hence, the major varieties were classified, investigated, and evaluated on the basis of three criteria: 1) number of operations required for solution, 2) rate of error accumulation, and 3) ease of computation of \bar{A}^{-1} . This part of the investigation is discussed in Section 2 below.

1.2 Solution of Singular or Poorly Conditioned Systems

It is well known that as the order of a square matrix (or the number m of columns in an $m \times n$ matrix, $m \leq n$) increases, the rank of the matrix (which cannot be greater than m in any case) has a tendency to decrease and the condition numbers to increase (where condition numbers are defined and have significance for rectangular matrices). These

tendencies result from computational and physical causes, not from mathematical causes. Increasing m means increasing the number of unknowns ΔX and, thereby, weakening the geometric structure which A describes. For example, each time a triangulation is barycentrically subdivided (which corresponds to increasing m of the matrix) the triangles which are elements of the subdivisions approach as a limit triangles with parallel sides. The same is true of elementary subdivisions, (but not of similarity -- preserving subdivision). Computationally, increase in m implies an increase in the number of computation operations involved. Each increase in number of operations increases the round-off error in the result.

Since near-singularity is a condition which would be expected frequently in large linear systems, we thought it worthwhile to investigate methods which would give answers whether the A matrix was poorly conditioned or even singular. A procedure for assuring that a solution of any square linear system was derived by J. Hawkins on purely theoretical grounds. A description of the procedure comparison with other procedures, and evaluation of the Hawkins procedure, is given in Section 3 below.

1.3 The Matrix as a Continuous Function

The properties of very small matrices can be worked out by considering each possible variation of the matrix elements. For matrices of order greater than, say, four, such an investigation procedure involves too many different possible cases, and more general rules of behavior must be found. So far, no complete description of those properties of a matrix that are important in solving the associated linear system has been achieved, in fact, knowledge on this subject is still sketchy. As order of the matrix

grows, however, the possibility arises of approximating the matrix by a function in two variables and of using the function properties as a description of the matrix properties. In particular, if:

$$A_{ij} \approx A(x,y)$$

then we look for the function N corresponding to $(A^T A)^{-1}$ and to $(A^T A)^{-1} A^T$. That function N then approximately describes \bar{A}^{-1} , has the same singularities and the same errors. The work done is described in Section 4 below.

1.4 The Matrix As a Geometric Object

Sets (of matrices) can be defined in such a manner as to be groups. Conversely, any group has a matrix representation. Further, groups can be identified with certain geometric surface properties so that there is a 1 to 1 correspondence between classes of groups and classes of surfaces. Hence, it seems reasonable to ask: What kind of classes of surfaces correspond to very large linear systems, and to what topological transformation does the inversion of a matrix correspond? If these questions can be found and are computationally meaningful, perhaps a study of surface properties will give valuable information about the properties of large matrices. Unfortunately, topology, which is the discipline concerned with surface properties that are invariant under continuous transformation, is a very recent development, is still largely intuitive in those areas which are significant, and is trivial elsewhere. One must, therefore, use care in using topology to trace the connection between matrices and surfaces. There is still very much to be done, and Section 5 below describes the problems involved in pushing further ahead, as well as describing what has been done.

1.5 Specific Computational Difficulties

There is no theoretical limit to the size of the system of linear equations that can be solved. There are many practical limits, however, and the most important of these are 1) the amount of time available for computations, 2) the number of significant figures that can be handled in a single operation, and 3) the rate at which round off errors accumulate. These three limitations are correlated to some extent; any solvable problem can be solved to any desired precision if enough time is allowed, for instance. One can nevertheless find a number of features which are common to all problems and which can be modified to improve the solution procedure for large linear systems. One of these features is the number of significant figures needed per number manipulated; another is relative positions of zero and non-zero elements. Section 6 discusses these matters.

2. REVIEW AND EVALUATION OF PROCEDURE FOR SOLUTION OF LINEAR SYSTEMS

2.1 Classification of Procedures

The sharpest distinction between solution procedures is afforded by using as criterion the number of steps needed to arrive at a satisfactory solution. A direct procedure is one in which the number N_0 of elementary operations (addition, subtraction, multiplication, division) needed to fix ΔX to a given number of significant figures is independent of the values of ΔY and A . An indirect procedure is one in which N_0 depends upon the values of ΔY and/or A . The indirect procedures can be further distinguished accordingly as initial approximate values of ΔX are or are not required, and accordingly as the number of operations can or cannot be shown to be finite. Some procedures, such as that of J. Hawkins (Section 3) can be shown to approach a solution regardless of the availability of approximate values for ΔX ; some of the gradient methods, on the other hand, require the approximate value of ΔX to be within a certain distance of the true value, if the sequence of operations is to give convincing results. (See Table 1)

These definitions are made on theoretical grounds only. They do not consider the way an actual computation procedure may be affected by the calculating machine used. For instance, although a large class of computational procedures can be shown to consist of convergent procedures for the particular problem:

$$A^T \Delta Y = A^T A \Delta X,$$

where $(A^T A)$ is positive definite for a real matrix A , the procedures may

CLASSIFICATION OF PROCEDURES

1. DIRECT PROCEDURES
 - 1.1 General
 - 1.1.1 Gauss (Reduction and Variants)
 - 1.1.2 Inversion
 - 1.1.3 Enlargement
Escalator and Bordering Methods
 - 1.2 Special
 - 1.2.1 Square Root Method
2. INDIRECT PROCEDURES
 - 2.1 Universally-Convergent Procedures
 - 2.2 Conditionally-Convergent Procedures
 - 2.2.1 Methods Involving Current Values Only
 - 2.2.2 Methods Using Previous Values

Table 1

actually diverge because of the accumulation of round-off errors. This is particularly true of iterative procedures where a poor first approximation may result in large correction numbers.

Suppose we have:

$$\begin{matrix} Y \\ n \times 1 \end{matrix} = \begin{matrix} A \\ n \times m \end{matrix} \begin{matrix} X \\ m \times 1 \end{matrix},$$

this can always be changed to a similar problem with a square matrix:

$$\begin{matrix} Y' \\ m \times 1 \end{matrix} = \begin{matrix} A' \\ m \times n \end{matrix} \begin{matrix} X \\ n \times 1 \end{matrix}$$

by the mappings

$$\begin{aligned} Y' &\equiv A^T y \\ A' &\equiv A^T A \end{aligned}$$

The matrix A' is in fact positive-definite and, in addition, symmetric. Any problem can, therefore, be converted to a problem with A' or A positive definite and symmetric. In most of what follows, however, it will only be assumed that A is square. This assumption does away with $\frac{mn(n+1)}{2}$ multiplications and $\frac{mn^2}{2}$ additions which are not really needed if A is already square.

It may be noted here that even conversion to a square matrix before solution is not necessary according to some procedures (Creusen, 1965). It can be shown, however, that these procedures are in fact mathematically the same as the presquaring procedures in that the $A^T A$ matrix is effectively formed during the solution process.

The class 1.1 of reduction methods starts from the complete set of linear equations:

$$Y = AX$$

where A' has all zeros below the main diagonal. The transpose is then taken,

$$(Y')^T = X^T (A')^T$$

$$(Y'')^T = X^T I$$

Mathematically and computationally, the reduction procedure can always be made to give a solution, regardless of the rank of A . This follows because if A is of rank r , the equations and unknowns can be re-ordered so that:

where A_{11} is non-singular. A general solution is the $\infty^{(n-r)}$ set consisting of arbitrary-valued X_2 and the unique set of values for X_1 gotten by solving:

It can be shown that the procedure described is the most efficient direct

procedure using only elementary operations for solving the general problem. This implies that it is also the most efficient direct procedure when the A matrix has certain special properties such as symmetry, antisymmetry, etc.

Table 2 shows (approximately) the number of operations and storage spaces needed for a direct solution of the general problem including the step from $m \times n$ to $n \times n$ matrix. If A is already square and symmetric, the number of operations required for $A^T A$ can be dropped. If it is merely square, however, this last number must be multiplied by some number between 1 and 3, depending on the type of solution employed.

Inversion procedures lie logically between reduction procedures, which step by step reduce the size of the system to be solved, and enlargement procedures which start from a small, known solution and build upwards from that point. In inversion procedures, the transformation from Y to X is done all at one time by first finding A^{-1} and then carrying out the operation:

$$A^{-1} Y = X \quad .$$

As mentioned previously, the inversion method as a means of finding X need not be considered seriously. First, it can be shown that in general the number of operations required cannot be less than the number required by the direct method; it will usually be greater. Second, there are many solution procedures which will give a solution regardless of the "condition" of the matrix. Because an inverse (or reciprocal) matrix has the reciprocal $\frac{1}{D}$ of the determinant of the original matrix as a factor of each element, the reciprocal matrix is not defined when the determinant D is zero.

DIRECT PROCEDURE

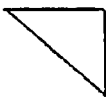
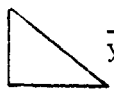
Memory		\pm	\times	\div	
A^T $m \times n$	A $n \times m$	$\frac{(2M+1)N+M^2}{2}$	$\frac{(N^2+N)(M-1)}{2}$	$\frac{(N^2+N)M}{2}$	0
$A^T A \rightarrow$		N^2	$\frac{2N^3-9N^2+13N-6}{6}$	$\frac{2N^3-9N^2+13N-6}{6}$	$\frac{N-N}{2}$
$A^T y$		$M+N$	$M(N-1)$	NM	0
$y \Rightarrow \bar{y}$			$\frac{(N-1)(N-2)}{2}$	$\frac{(N-1)(N-2)}{2}$	N
$x =$	 \bar{y}	N	$\frac{N^2-N}{2}$	$\frac{N^2-N}{2}$	0
TOTAL			$\frac{(N^2+3N-2)M}{2}$ $+ \frac{2N^3-6N^2-2N}{6}$	$\frac{(N^2+3N)M}{2}$ $+ \frac{2N^3-3N^2+N}{6}$	$\frac{N^2+N}{2}$

Table 2

Nevertheless, reciprocal matrices should be considered for their own sake, since they are important in probability theory and will in all practical cases have to be computed regardless of whether or not they are used in the solution. When reduction procedures are used for the solution, computation of the reciprocal matrix can conveniently be carried along at the same time and by the same procedures. This follows because from the definition of a reciprocal matrix:

$$I = AA^{-1} \quad ,$$

we see that I and A^{-1} can be broken up into column vectors:

$$I_j = AA_j^{-1} \quad .$$

The elements of A_j^{-1} are solved for just as were the elements of X . Naturally, if exactly the same procedure were used there would be n times as many operations required as in solving for X , since there are n columns A_j^{-1} ; this would be about $\frac{n^4}{3}$ operations. Since only one element of I_j is different from zero, however, only those operations need be made which relate to the non-zero element, and it can be shown that the number of operations actually increases only as (again approximately) $\frac{n^3}{3}$. Hence, simultaneous computation of X and A^{-1} can be done efficiently, especially if computations of the two are interlaced, i.e. if the procedure is planned to have quantities computed in one part used in another when possible. Table 3 gives estimates of the computational characteristics of the (direct) reduction method and of the (direct) biorthogonalization method for inverting matrices, and includes for comparison the characteristics of these methods and of the iteration method for solving the main problem. Other inversion procedures will be discussed along

MATRIX INVERSION
Operation and Storage Counts

Method & Prob.	Add. & Subt. ¹	Multi. ¹	Division ¹	Min. Stg.
Reduction Y = AX	$\frac{n}{6} (n-1) (2n+5)$ each		$\frac{1}{2}n (n+1)$	$n(n+1)$
Reduction A^{-1}	$\frac{n}{6} (n-1) (8n-1)$ each		$\frac{1}{2}n (3n-1)$	$2n^2$
Reduction A^{-1} (Sym. A)	$\frac{n}{6} (n-1) (2n-1)$	$\frac{n}{3} (n^2-1)$	$\frac{1}{2}n (n+1)$	$\frac{1}{2}n (n+1)$
Biorthogonal A^{-1}	$2n^2 (n-1)$	$2n^3 - n^2$	n^2	$2n^2$
Iterative Y=AX per cycle	$n (n+1)$	$n (n+1)$	0	DATA + n
Escalator A^{-1}	$4n^3 / 3$	$2n^3 / 3$ Total		n^2
Matrix Modif. A^{-1}	$2n^3$	$2n^3$		n^2+2n
Iteration A^{-1} (totaling) per cycle	$3n^3 + 2n$	$3n^3$	0	$3n^2$

¹ The number of divisions can be reduced to n in every case by storage of approximate reciprocals. The number of multiplications is increased accordingly.

Table 3

with the discussions of the procedures for solving the main problem.

The square root or Banachiewicz method (Banachiewicz, 1938) is one of a class of direct methods designed to take advantage of special matrix characteristics -- in this case, symmetry. The symmetric matrix A is split up into the product of two matrices:

$$A \equiv BTB$$

where $b_{ij} = 0$ for $i > j$. Just as in the Gauss procedure, a vector K is found from

$$Y = B^TK \quad ,$$

and X is found from

$$K = BX$$

by the same method. Although twice as many steps are required here as in the corresponding Gauss procedure, the first part, in which B is computed, makes up at least partly for this. Here:

$$b_{ii} = a_{ii} - \sum_{k=1}^{i-1} b_{ki}^2$$

$$b_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} b_{ki} b_{kj}}{b_{ii}} \quad j \geq i$$

$$b_{ij} = 0 \quad j < i$$

A careful analysis of the procedure shows that the number of operations is actually somewhat larger than the number of an efficient Gauss procedure. This is especially true if the number of square root operations is turned

into an equivalent (for the computer) number of multiplications and additions. Computation of a square root requires about $150\mu\text{s}$ on an IBM 7094, while a multiplication takes about $5\mu\text{s}$. Since only n square roots are involved, however, this difference is insignificant when large systems are to be solved. Table 4 shows the estimated number of operations involved.

SQUARE ROOT PROCEDURE

<u>Add. and Subt.</u>	<u>Multiplication</u>	<u>Division</u>	<u>Sq. Rt.</u>
$1/3 [n^3+5n^2-4n+3]$	$1/3 [n^3+5n^2-4n+3]$	$1/2 (n^2+5n-2)$	n

Table 4

2.2 Partition Procedures

Partition procedures could be classed as reduction procedures since they involve breaking the main system of equations into a number of smaller systems each of which is more easily solvable (or has been solved). It is included among the inversion methods because the subsystems selected are independent, and non-overlapping, whereas the subsystems computed in reduction or enlargement procedures are subsets of each other (in sequence).

As with many other procedures, we find that the Frobenius-Schur lemma (Bodewig, 1959) is the basis for the procedure. The system:

$$Y = AX$$

is partitioned into subsystems:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} ,$$

such that A_{11}^{-1} , and

$$\Delta^{-1} \equiv (A_{22} - A_{21} A_{11}^{-1} A_{12})^{-1}$$

can be easily computed. Then:

$$A^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

is composed of elements

$$B_{11} = A_{11}^{-1} (I + A_{12} \Delta^{-1} A_{21} A_{11}^{-1})$$

$$B_{12} = A_{11}^{-1} A_{12} \Delta^{-1}$$

$$B_{21} = \Delta^{-1} A_{21} A_{11}^{-1}$$

$$B_{22} = \Delta^{-1} .$$

(The Frobenius-Schur lemma is a special case of a lemma applicable to partitioning into an arbitrary number of subsystems:

$$[Y_i] = [A_{ij}] [X_j] .$$

It can be derived by judicious application of Cramer's rule.)

Evidently partition procedures offer no particular saving in time or storage space. They are of importance when either some of the subsystems have already been solved or (what is almost the same thing) when the partitioning can be carried out so as to make enough of the

A_{ij} into zero matrices that computation time is saved. In all subsequent discussions it will be assumed that the main system has already been so partitioned, and that the system under discussion cannot be usefully partitioned further.

2.3 Enlargement Procedures

Under the enlargement category of procedures are grouped those procedures which solve a sequence of linear systems each of which is a proper subset of the succeeding system. They are distinguished in this way from the reduction procedures in which the sequence runs the other way, the size of the system to be solved being whittled down eventually to one equation in one unknown, after which, by successive substitutions, the rest of the system is solved as a sequence of single equations in one unknown.

Enlargement procedures are, along with partition procedures, based on the Frobenius-Schur lemma:

$$A^{-1} \equiv \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} A_{11}^{-1} A_{12} \Delta^{-1} A_{21} A_{11}^{-1} & - A_{11}^{-1} A_{12} \Delta^{-1} \\ - \Delta^{-1} A_{21} A_{11}^{-1} & \Delta^{-1} \end{bmatrix}$$

where,

$$\Delta^{-1} \equiv [A_{22} - A_{21} A_{11}^{-1} A_{12}]$$

The linear equation problem:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = [A] \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \end{bmatrix},$$

then has the solution:

$$\begin{bmatrix} \Delta X_1 \\ \bar{X}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & \Delta^{-1} (A_{21} X_1 - Y_2) \\ -\Delta^{-1} (A_{21} X_1 - Y_2) \end{bmatrix}$$

where

$$\Delta X_1 = \bar{X}_1 - X_1$$

and X_1 is a solution of

$$Y_1 = A_{11} X_1.$$

The above procedure is particularly simple when A_{11}^{-1} and X_1 are already known and when A_{22} is a scalar; in other words, when X_2 consists of a single element, for in that case Δ is a scalar, its reciprocal is immediately computable, and the formula for

$$\begin{bmatrix} \Delta X_1 \\ \bar{X}_2 \end{bmatrix}$$

can be rapidly evaluated.

This same procedure gives, of course, also the reciprocal matrix, at the cost of additional storage space and a little bit more of computation. There are numerous variants of this procedure -- the bordering and escalator procedures are best known. Sequential analysis is also related to the above procedure, since we have:

$$\Delta X_1 = \Delta^{-1} \{A_{11}^{-1} A_{21}^T (A_{21} X_1 - Y_2)\}$$

where

$$\Delta^{-1} \equiv (I - A_{11}^{-1} A_{21}^T A_{21})^{-1}$$

and the original problem is

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} X_1 + \Delta X_1$$

where

$$Y_1 = A_{11} X_1 \quad .$$

There is obviously little difference between reduction procedures and enlargement procedures as far as number of operations required is concerned. Major differences occur in the amount of storage space needed and in the types of numerical checks (controls) used on the computations. There is also a very considerable difference in the conditions under which the procedures can be used; reduction procedures must apparently start with the complete system, whereas the enlargement procedures can start with whatever data are available and step-by-step build up the final solution as more data arrive and more unknowns are added. This difference is more apparent than real, for reduction procedures can be written to compute downward from a given system to the previous smaller system.

It is worth pointing out here that the enlargement methods are similar to the iterative indirect methods to be discussed below. Both sets of procedures start from assumed solutions and work towards a final solution; for enlargement procedures, the assumed solution happens to be also an exact solution to a subset of the entire system. Because of this similarity, the enlargement and iterative indirect procedures can easily be used together on the same problem, or techniques useful to one set of procedures can be applied to the other. This fact has many interesting results which will be discussed in greater detail in a future report.

2.4 Orthogonalization Procedures

Orthogonalization procedures have features linking them to both direct and to the indirect procedures. They are certainly direct procedures, since the number of steps involved is independent of the values of A , Y , or the starting value X_0 . On the other hand, they start from an assumed value X_0 of X , which makes them akin to the indirect methods. The final value, X , is built up from X_0 in a series of steps such that:

$$X = X_0 + \sum_i k_i \Delta X_i ,$$

where the k_i are coefficients and the ΔX_i are orthogonal with respect to some preselected matrix. (Note the similarity of the procedure to expansion of a function into orthogonal polynomials). Because of the orthogonality of the ΔX_i one can easily show that 1)

$$\left\| \sum_j^n k_j \Delta X_j \right\| \geq \left\| \sum_{j+1}^n k_j \Delta X_j \right\|$$

and that 2)

$$k_i \Delta X_i = 0 \text{ for } i > n$$

where n is the rank of A .

The conjugate gradient method is an orthogonalization procedure in which the ΔX_i and k_i are derived from the successive residuals ΔY_i .

$$Y_i = Y - Y_i$$

$$Y_i \equiv AX_i$$

$$X_i \equiv X_0 + \sum_{j=1}^i k_j \Delta X_j$$

Its computational characteristics are shown in Table 5.

CONJUGATE GRADIENT PROCEDURES

	<u>Storage</u>	<u>Add. & Subt.</u>	<u>Mult.</u>	<u>Div.</u>
Single Step	$6n^2+18n+7$	$2n^2+5n-4$	$3n^2+7n$	2
Cycle	$6n^2+18n+7$	$n(2n^2+5n-4)$	$n(3n^2+7n)$	2n

Table 5

Several characteristics of orthogonalization procedures are of importance.

- 1) Although the procedures are in theory applicable to any non-singular system, they are practically limited to systems with positive definite matrices. Solution of a general system involves at some stage conversion to a positive definite system, introducing $n^3/2$ multiplications and $n^3/2$ additions (approximately).
- 2) The number of operations in an orthogonalization procedure is greater than the number in a reduction direct procedure by a factor of at least 3 in the general case. Its use must, therefore, be justified on grounds other than that of economy of operation. Justification follows from the fact that the smaller the difference $X - X_0$, the smaller are the successive corrections kX_i and, therefore, it can be assumed that the computation sequence can be set up to reduce round-off error accumulation below what is encountered in the reduction direct procedures. It seems reasonable to expect that round-off error accumulation could be minimized by setting:

$$X_0 + \sum_{j=1}^n k_j \Delta X_j = X_0$$

and starting a new cycle in an indirect procedure such as the Gauss-Seidel.

2.5 Indirect Procedures

2.5.1 Universally-Convergent Procedures

To retain the precision of the classification given in Table 1, a distinction is made between those procedures which always approach a limit regardless of the make-up of the system and those procedures whose convergence depends on characteristics of the system, on the initial solution assumed, etc. A further distinction could be made between n-step procedures, which theoretically are complete after n-steps (where n is the order of the A matrix), and unlimited step procedures which have no pre-determinable number, of steps to convergence. There are many points of similarity between the reduction direct procedures and n-step methods that these are usefully placed in the equivalence (inversion) group.

As an example of the universally-convergent procedure the projective method of J. Hawkins (1967) may be cited. This procedure is more fully described in the next section, but in simplest terms it defines the solution of:

$$Y = AX$$

to be the coordinates of intersection of a certain plane with the perpendicular from the origin to that plane, and solves for a sequence of planes and foot-points. There are many resemblances between the Hawkins, Kacmarz, and Cimmino procedures, so it may be assumed that the number of operations involved are of the same order of magnitude for each. Table 6 gives estimates of the computational effect involved.

UNIVERSAL PROCEDURES

	<u>Storage</u>	<u>Add. & Subt.</u>	<u>Multiplication</u>	<u>Division</u>	<u>Sq. Rt.</u>
Kacmarz	$n(n+3)$	$n(3n-1)$	$3n^2$	$n+1$	n
Cimmino	$n(n+3)$	$(3n^2-2n+1)$	$3n(n+1)$	$(n+1)$	n
Hawkins	$n(n+2)$	$n(3n-1)$	$2n(n-1)$	$n-1$	0

NOTE: In above procedures, change 3 to 2 for iterations subsequent to first.

Table 6

2.5.2 Conditionally Convergent Procedures

A conditionally convergent procedure is one whose convergence depends upon the numerical values of the A matrix elements. The universally convergent procedures discussed above and in Section 3 will converge at least to some value of X regardless of the rank of A or its condition numbers. For example, the commonly used Gauss-Seidel and over-relaxation methods require for convergence that A be positive definite; this is a severe restriction. The advantages of conditionally convergent procedures are that:

- 1) although the number of steps required to arrive at an exact answer is in theory infinite (in general), the number of steps needed to reduce the deviation from the exact answer to a tolerable value is finite and may be less than the number required by a direct procedure.
- 2) The precision of an indirect procedure is limited to round-off error accumulation. This error limit is, however, approximately proportional to $n^{2/3}$ and is independent of the number of steps taken. The round-off error limit for any direct method is approximately proportional to $n^{4/3}$. If, for example, 4 significant figures are lost in the indirect method because of round-off, 8 significant figures could be lost solving the same problem with a direct procedure.
- 3) As a general rule, the amount of storage space (in bits) needed by direct methods is not (except in certain variants of the orthogonalization procedures) lessened by any prior knowledge. Indirect procedures, in general, are benefited by using prior knowledge -- the number of steps and amount of storage space needed are both diminished.

Table 7 shows the minimal computational requirements for any iterative procedure for the first cycle and for the total number of cycles and steps needed to bring all elements of X to the required precision. Since (assuming $m=n$) the number of operations needed by a direct procedure cannot be less than about $n^3/3$, a conditionally convergent procedure can go through $n/3$ complete cycles before losing to direct procedures inefficiency. Furthermore, at this point it will have lost only half as many significant figures through round-off error.

The advantages of indirect procedures are balanced or cancelled by several other considerations.

- 1) If, as is usually the case in practical problems, the reciprocal matrix must be computed or $A^T A$ computed, the number of operations for the combined computation is immediately of the order n^3 regardless of whether or not indirect procedures are used.
- 2) Convergence for many indirect procedures is guaranteed only for special kinds of matrices -- usually positive definite matrices. The rate of convergence is dependent on the X_0 initially chosen and to be made reasonably large may further require knowledge (usually by computation) of quantities not directly accessible such as the A matrix spectral radius.

Table 8 shows the computer requirements of two common convergence-conditional (C.C.) indirect procedures; the point relaxation and gradient procedures. All of the procedures listed in the table could be classified as relaxation procedures, since they are variants of the same equation:

$$X^{(m+1)} = BX^{(m)} + CY \quad .$$

ITERATIVE PROCEDURES

Storage		Addition		Multi.
Ax	$n(2m+1)$	$n(m-1)$		nm
y-Ax	2n	n		
TOTAL (per cycle)	$2nm+3n$	mn		nm
TOTAL c = number of cycles k = fraction of cycle	$2nm+3n$	$(c+k)mn$		$(c+k)mn$

The smallest possible procedure involves $\bar{y} = Ax; y = \bar{y}$.

Table 7

C.C. INDIRECT PROCEDURES

Relaxation	<u>Storage</u>	<u>Add. & Subt.</u>	<u>Mult.</u>	<u>Div.</u>
1) Gauss-Seidel	n^2+2n	n^2	n^2-n	n
2) Over, Under	n^2+2n+2	n^2+1	n^2-n+2	n
Steepest Descent	n^2+4n+1	$2n^2+2n-2$	$2n^2+3n$	1

Table 8

The equation:

$$X^{(m+1)} = (D + \omega L)^{-1} \{ (1-\omega)D - \omega U \} X^{(m)} + \omega Y .$$

represents in fact, the procedures:

$$\begin{array}{l} \text{under-relaxation} \\ \text{Gauss-Seidel} \\ \text{over-relaxation} \end{array} \quad \text{according as } \omega \left\{ \begin{array}{l} < \\ = \\ > \end{array} \right\} 1 .$$

U, D, and L are the upper triangular, diagonal, and lower triangular components of A.

C.C. Indirect Procedures are even more varied than the direct methods, and there are interesting relations between them. The relation of the most important of these to very large systems will be studied in more detail in Section 6.. Some comments at this point are necessary.

- 1) Most procedures of this type require that the matrix A be positive definite. The reason is geometrically obvious. Successive approximations take one inward toward the center of a family of surfaces. A positive definite matrix defines a family of hyper-ellipsoids, and there is no difficulty in proceeding to the center. If the matrix is not positive definite, the hyper-surfaces (corresponding, e.g. to "hyper"-hyperboloids) may diverge to infinity and successive approximation diverge along with them. This difficulty could be overcome, but possibly only by modifying the procedure to the point where it is no longer competitive with a direct procedure.
- 2) The estimates given in Table 8 for over and under relaxation methods do not include the operations needed for computation of ω . If ω is computed for maximum rate of convergence the value of ω which minimizes the spectral radius (i.e. maximum eigenvalue) of

$$(D+\omega L)^{-1} \{ (1-\omega)D - \omega U \} \quad ,$$

must be found. The process of finding or estimating the spectral radius and its minimizing parameter ω can be simple or complicated.

3. THE HAWKINS UNIVERSALLY CONVERGENT INDIRECT PROCEDURE

The Hawkins' procedure was described in a previous report, CR-66-198-1. It was derived to have available a procedure that would converge under all conditions, but extension of the principles used in it to other C.C. indirect procedures has seemed to have promise. Further, investigation of the method reveals that it has many points of resemblance to the methods of Kacmarz and of Cimmino (Bodewig, 1959). There are differences, however, which could be valuable.

Suppose our linear system (of dimension n) is:

$$Ax = f \quad . \quad (1)$$

Kacmarz's method is given by:

$$x_{\ell+1} = x_{\ell} + \alpha_{\ell} A_i$$

where

$$\alpha_{\ell} = - (x_{\ell} A_i - f_i) / A_i^2 \quad .$$

Usually, however, A is transformed beforehand so that $A_i^2 = 1$, in which case:

$$\alpha_{\ell} = - (x_{\ell} A_i - f_i) \quad .$$

To prepare system (1) for my method, first divide each equation by f_i , so that:

$$A'_i x = 1,$$

where

$$A'_i = A_i / f_i \quad .$$

In matrix notation:

$$A' x = f' = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} . \quad (2)$$

The row vectors of A' define an $n-1$ dimensional hyper-plane H . If P is the point in H whose associated vector (from the origin) is normal to H , then:

$$x = P/P^2 .$$

Given a point x_ℓ in H and a row A'_i of A' , the method is given by:

$$x_{\ell+1} = x_\ell + \alpha_\ell (x_\ell - A'_i)$$

where

$$\alpha_\ell = \frac{-x_\ell (x_\ell - A'_i)}{(x_\ell - A'_i)^2} .$$

Where Kaczmarz iterates with the same set of vectors (the A_i), the above method creates a new vector $(x_\ell - A_i)$ for each iteration. The formula for α_ℓ is essentially the same in each case.

Suppose from system (2) we form $n-1$ vectors:

$$V_i = A_i - A_1, i=2, \dots, n \quad (3)$$

then the iterative process becomes:

$$x_{\ell+1} = x_\ell + \alpha_\ell V_i$$

where

$$\alpha_\ell = -x_\ell V_i / V_i^2 .$$

The process in this case is exactly the same as that of Kaczmarz.

The difference mentioned earlier is that now we are dealing with $n-1$ vectors (the V_i) instead of n (the A_i). Thus, the rate of convergence depends on the $n-1$ vectors V_i . Suppose we rewrite equation (3),

$$V_i = \lambda_i A_i - A_1 \quad . \quad (3)'$$

The condition of the vectors V_i now depend on the $n-1$ real numbers λ_i . Hopefully, a simple iterative technique can be found which produces the optimum set of λ_i , i.e. the λ_i which optimize the condition of the V_i . In Table 8 (Section 2.5.2 preceding) is given a comparison of those characteristics of the Kacmarz, Cimmino, and Hawkins procedures which are of importance computationally.

4. THE MATRIX AS A CONTINUOUS PROCESS

In the preliminary report CR-66-198-1, a description was given of the investigations made into densification procedures. The reasoning is that as the size of linear systems

$$\Delta Y = A \Delta X$$

increases (i.e. as the rank of A increases) there is a possibility that some of the gross characteristics of the system can be approximated by the analogy:

$$y(i) = \int_i \int_j a(i,j) dx(j).$$

There is no mathematical justification for this assumption, but there is adequate physical justification. The physical meaning of allowing n to increase is that a very large number of samples (observations) are taken from the universe described by A and that the number of observations and the number of causative factors keep in step. Since this would imply the impossibility of adequately describing the universe by means of a small number of variables, physics requires that the X vector be describable as a function of a finite, usually small, number of variables. For examples, the Y vector might be a series of range or angle measurements and X the corresponding positions of a space probe or satellite; Y might be the measured coordinates of star images and X the corresponding set of star positions; Y might be direction and height measurements of ground points and X the spheroidocentric coordinates of these points; etc., etc.

In every example except perhaps the second, it is obvious that the X vector is not composed of completely independent variables. Even in the astrometric example, the elements of X are connected by common membership in a globular cluster or in the same association, or in the same galaxy, etc. Therefore, one can justify the assumption that when n is increased far enough relations begin to appear between the elements of X.

Nevertheless, there are serious objections to use of this analogy, and investigation into this aspect is going on. A separate report will be made on the results of these investigations.

5. INVESTIGATION OF TOPOLOGY OF LARGE MATRICES

In the INTRODUCTION a connection was traced (in general terms) between matrix theory and topology theory. (The term "theory" is used here in the same sense as defined in Merriam-Webster Second International Dictionary. It will never be used to designate a particular kind of set as is sometimes done in topology theory.) The connection is possible because parts of each theory can be described in terms of group theory concepts. It can be made more concrete as follows.

Let A be an $n_0 \times m_0$ matrix; suppose $n_0 \geq m_0$ with the understanding that results for $n_0 \geq m_0$ apply with appropriate interchange of terms of matrices for which $n_0 \leq m_0$. By addition first of row multiples and then of column multiples A is converted to the canonical form:

$$\begin{bmatrix} A & 0 \\ B & 0 \end{bmatrix} = \left[\begin{array}{ccccccc} & \epsilon_{11} & & & & & . \\ & & 0 & & & & . \\ & \epsilon_{22} & & & & & . \\ & & . & & & & . \\ & & & . & & & . \\ & & & & . & & . \\ & & & & & \epsilon_{ii} & . \\ . & . & . & . & . & . & . \\ & b_{ij} & & & & & . \\ & & & & & & . \end{array} \right] \quad \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ 0 \end{array}$$

where:

ϵ_{ii} is + 1 or 0,
 b_{ij} is 0 if ϵ_{ii} is 0, and
+ 1 or 0 otherwise.

The following statements are then self-evident.

Any matrix Q ($n \times m_0$) with

$$q_{ij} = 0 \quad \text{for } j < m_0 - r, \quad \text{and} \\ \text{for } j = k \text{ if } \varepsilon_{kk} = 0$$

can then be written as a multiple of

$$\begin{bmatrix} A & 0 \\ r \times m_0 \end{bmatrix},$$

$$\begin{aligned} Q &= R \quad [A \ 0] \\ n \times m_0 &= n \times r \quad r \times m_0 \end{aligned}$$

or, if $\begin{bmatrix} \bar{A} & 0 \end{bmatrix}$ is that submatrix of $\begin{bmatrix} A & 0 \end{bmatrix}$ containing only the non-zero rows,

$$Q = R \begin{bmatrix} \bar{A} & 0 \end{bmatrix}$$

For many purposes the matrices $[Q]$ and $[\bar{Q}]$, where $[\bar{Q}]$ is the submatrix of $[Q]$ with the columns $m > r$ omitted, mean the same thing, in which case Q is generalized to:

$$Q(n \times m; m \leq m_0) \quad .$$

Furthermore, if the transformation is limited to the type

$$A' \equiv T^{-1} A T$$

where T is a non-singular matrix, then A can be transformed into the Jordan cononical form:

A diagram showing a triangular arrangement of points. The points are arranged in rows, with the first row having one point labeled A_1 , the second row having two points labeled A_2 and \cdot , and the r -th row having r points, the last of which is labeled A_r . To the right of the points in the second and r -th rows are the labels 0 . A horizontal line of points is shown below the main triangle, with the first and last points labeled 0 .

where the A_i are of the form:

$$\begin{bmatrix} \lambda_i & 1 & 0 & & \\ 0 & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & 0 & & & \lambda_i \end{bmatrix},$$

with P_i eigenvalues along the main diagonal, 1's along the $A_{j,j+1}$ sub-diagonal, and zeros everywhere else. If A is the given matrix, then the number of zero columns in the Jordan canonical form is the Betti number of the group generated by A ; the Torsion coefficients are the products;

$$\pi P_i \quad i = 1, 2, \dots, n.$$

It can easily be shown that two matrices are similar if, and only if, they have the same Betti numbers and torsion coefficients. Now a finitely-generated commutative group can obviously be represented as the product of 1) a number of groups which are not cyclic, and 2) a number of sub-groups which are cyclic. The number of non-cyclic sub-groups required is the Betti number; the orders of the cyclic groups are the torsion coefficients, and the finitely-generated group is determined to within an isomorphism by the Betti-numbers and torsion coefficients. The "obviousness" follows if the A matrix and the matrices derived from it by similarity transformations are regarded as group representations.

Relating surfaces and groups is slightly more difficult than relating matrices and groups. Much of the difficulty in topology theory

arises from 1) the lack of rigor in the theory, and 2) from semantic difficulties caused by attempting to have the theory include various kinds of non-finite sets. Careful analysis of the logical structure of topology theory shows, however, that theory is reducible to a few concepts of real importance.

- 1) It is assumed that any surface of interest can be approximated well enough by a set of connected triangular surfaces.
- 2) Defining point and line to be degenerate forms of a triangulation can be extended to any number of dimensions.
- 3) The subject matter of topology is limited by implicit definition to those properties of surfaces which can be described by 1) a simple listing of the vertices of the associated triangulation and 2) the rules for connecting the vertices and triangular surfaces. In fact, the term "surface" in topology means simply a) a set of points (in the Euclidian sense) that are related to each other by the given allowable rules or b) a set of quantities (geometric or mathematical) which can be related to the point set (1) in such a way that the rules for (1) are not violated. Combinational topology theory (the theory of surfaces as it can be described as combinations of points) for esthetic reasons starts at a level of generality far greater than can be accommodated by the inevitable restriction that the surface must be describable as a set of symbols P_i with rules connecting the symbols.

In view of the preceding general remarks discussion of the surface-to-group relationship will be carried out at the intuitive level; rigorous derivation of non-trivial results would require among other things a revision of the basic theories of combinational and set topology, and this is far outside the scope of the present investigation. Furthermore,

although the discussion will be in terms of triangular surfaces, the discussion can apply to any dimension by substitution of suitable terms.

- 1) A surface is a set of triangular surfaces (TS) which are combined according to certain rules. These in essence give meaning to linear combinations of TS, and such combinations are called chains and denoted by C^k where k is the (Euclidean) dimension number of the surfaces. That is, a chain C^k is:

$$C_i^k = \sum_{j=1}^n a_{ij} S_j^k$$

where i denotes a particular chain and S_j^k is a k dimensional element of the surface.

- 2) The chains C^k which "describe" a particular surface form a group. They, therefore, have bases. Furthermore, each base C^k has a boundary chain associated with it:

$$\begin{aligned} C^{k-1} &\equiv \partial C^k \\ &= \sum_v \sum_j a_{ij} \epsilon_v^{kj} s_v^{k-1} \\ &\equiv \sum_v b_{iv} s_v^{k-1} \end{aligned}$$

where: $\epsilon_v^{kj} = \begin{cases} +1 \\ 0 \\ -1 \end{cases}$ accordingly as,

s_v^{k-1} are $\left\{ \begin{array}{l} \text{incident with same orientation} \\ \text{not incident} \\ \text{incident with opposite orientation.} \end{array} \right\}$.

Hence, a surface generates a sequence of matrices Q^k whose columns are labelled by the elements of the k base chain and the rows by the elements of the base of the $k-1$ boundary chain, and whose elements are ϵ_v^{kj} .

-

At this point, therefore, we have established the sought-for relation between surfaces and matrices, and this connection will for the time being suffice. The next step is to relate the computability of a large system to the surfaces characterized by the torsion coefficients and Betti numbers, to see if topology theory can assist in simplifying the computational procedure. A beginning on this has been made. It is described in Section 2 of the preceding report. In that report, the quantities C'_{ij} , C^C_{ij} were defined in such a way that they gave the amount of incidence between rows i and \bar{j} or between columns i and \bar{j} , respectively. Their utility in estimating the computability of a system is shown for the computation of:

in Table 9. The relations between (C_{ij}^r, C_{ij}^c) torsion coefficients τ^k and Betti numbers β^k will be investigated to see:

COMPUTATION OF $A^T A$

METHOD	MEMORY IN/OUT	\pm	\times
1. $C_{ij}^r, C_{ij}^c \neq 0$	$\frac{(N^2+N)(2M+1)}{2}$	$\frac{(N^2+N)(M-1)}{2}$	$\frac{(N^2+N)M}{2}$
2. $C_{ij}^c \approx 0$	$\frac{(N^2+N)(2\bar{C}_{ij}^c+1)}{2}$	$\frac{(N^2+N)(\bar{C}_{ij}^c-1)}{2}$	$\frac{(N^2+N)\bar{C}_{ij}^c}{2}$
3. $\bar{C}_{ij}^r \approx 0$	$\frac{(N\bar{C}_{ij}^c+N)(2M+1)}{2}$	$\frac{(N\bar{C}_{ij}^r+N)(M-1)}{2}$	$\frac{(N\bar{C}_{ij}^r+N)M}{2}$
4. $\bar{C}_{ij}^r, \bar{C}_{ij}^c \approx 0$	$\frac{N(\bar{C}_{ij}^r+1)(2\bar{C}_{ij}^c+1)}{2}$	$\frac{N(\bar{C}_{ij}^r+1)(\bar{C}_{ij}^c-1)}{2}$	$\frac{N(\bar{C}_{ij}^r+1)\bar{C}_{ij}^c}{2}$

Table 9

- 1) how far one set may be computed from the other;
- 2) how these quantities change with increasing size of the system; and
- 3) what their geometric significance is as the size of the system increases and the matrix is densified.

It appears at present as if the (C_{ij}^r, C_{ij}^c) is easier to use than the (τ^k, β^k) set and has more computational significance (or at least is easier to interpret computationally).

6. COMPUTATION PROCEDURES

In the previous report there was a discussion of the effectiveness of various kinds of iterative (i.e. indirect) procedures for solving very large systems. The conclusion was reached that the only indirect procedure which would function efficiently in general was one which 1) started off from a reasonably close approximate solution, 2) converged regardless of the "condition" of the A matrix of the system, and 3) made use of the results of preceding iterations to improve the results of the current iteration. Many details are gone into in the previous report and are not needed here. An outline of the "optimum" procedure, as it looks at present, is given below to show the relation of the various parts of this report to the over-all investigation.

6.1 The Linear System

$$Y = AX$$

is to be solved for X and for $(A^T A)^{-1}$.

6.2 Preliminary Handling and Computation

6.2.1 The system is approximated by continuous functions if feasible and approximate values of X and $(A^T A)^{-1}$ computed, along with such estimates of the errors in X and $(A^T A)^{-1}$ as may easily be computed. The dN error in

$$N \equiv (A^T A)^{-1}$$

that is caused by errors in $(A^T A) \equiv n$ is well known; it is

$$dN_{ij} = - \sum_k \sum_\ell N_{ik} N_{j\ell} dn_{k\ell} \quad .$$

To compute [dN] by matrix procedures would be not only difficult but foolish at this stage. Computation by approximate methods may be feasible.

6.2.2 If the results of 6.2.1 are favorable, the system is rearranged to group zero elements and non-zero elements into blocks, etc.

6.2.3 Topological invariants on the C_{ij}^r , C_{ij}^c numbers are computed.

6.3 The matrix $A^T A \equiv n$ is computed.

6.4 Using an approximate value of X derived from 6.2.1 or from an approximate solution of

$$A^T y = \bar{n} X$$

where \bar{n} is a matrix derived from n by suppression of elements according to a scheme dictated by the numbers derived in 6.2.2, an iterative procedure is started.

$$\begin{aligned} X^{(m+1)} &= X_0 + \delta X^{(m)} \\ \delta X^{(m)} &= A^m \delta X^{(m-1)} + B^m Y \\ N^{(m+1)} &= N_0 + \delta N^{(m)} \\ \delta N^{(m)} &= C \delta N^{(m-1)} \end{aligned}$$

The successive vectors $X^{(m)}$ will contain fewer and fewer significant figures or will contain a constant number with decreasing maximum magnitude. The $\delta X^{(m)}$ form a sequence whose value at step (m+r) is estimated. The iteration then proceeds from $\delta X^{(m+1)}$ directly to $\delta X^{(m+k)}$. As m increases, the value of k can be expected to increase also.

6.5 Computation is stopped when a preset limit on some function of δX and/or δN is reached.

7. SUMMARY

Investigation of procedures for solving very large linear systems has gotten to the following point.

7.1 A procedure has been derived for solving a linear system for any order of any rank, and with arbitrary sized condition numbers. This procedure may be considered a variant of the Kacmarz and Cimmino procedures, which has some advantages in computation characteristics and in flexibility.

7.2 A start has been made on the simulation of very large systems by functions. Such simulation is inadequate for exact computation but may have advantages for 1) deriving initial approximate solutions rapidly, and 2) estimating the effects of errors in the Y and A matrices on the X matrix.

7.3 Tentative identification has been made of computability criteria, the C_{ij}^r and C_{ij}^c numbers, which can be used in planning the solution of very large linear systems. These numbers have other properties which relate them to incidence matrices and hence to geometric structures which can be investigated by topology theory.

7.4 A study of the computational characteristics of various known procedures shows that for solution of very large linear systems an iterative procedure is probably the best. It is suggested that the most flexible procedure, which at the same time is reasonably efficient, would be one involving the following steps.

7.4.1 An initial approximate solution is obtained by a direct method or by replacing the system by a bivariate function.

7.4.2.1 If the system is singular or poorly conditioned, a universally convergent procedure modified to allow use of production sub-procedures may be required.

7.4.2.2 If the system is reasonably well conditioned, a conditionally convergent indirect (iterative) procedure adapted to the particular system being studied should be used. Methods of adaptation are still being investigated; they may be related to topological invariants discussed earlier. Predictive procedures (also called semi-iterative or universal procedures) are apparently superior to or at least as good as other procedures.

7.4.3 Where needed, the reciprocal matrix of the system is computed if possible at the same time as the solution. Otherwise relevant portions of the solution are stored (on magnetic tape) and used in a later computation of the reciprocal matrix.

7.5 Direct sequential procedures are inferior to straight Gaussian procedures or indirect procedures except in those cases where the system to be solved is not all present at one time but is presented in parts, as in the immediate reduction of continually arriving data. Even here, the direct sequential procedures may be inferior to indirect procedures if the ultimate in precision is wanted.

7.6 Within the limits of the amount of time left, further investigation is planned to procede along the following lines.

7.6.1 Study of simulation of a large system by bivariate functions will be pushed. Special attention will be given to use of such functions for computing an approximate solution of the system and for computing error estimates.

7.6.2 The relation of the computational characteristics of very large linear systems to topological invariants will be pursued.

7.6.3 Further equations will be derived for use in predictive indirect procedures, and the solution process for very large linear systems formalized.

7.6.4 Little attention has been paid so far to the propagation of errors through the system. This will be given close attention in the next phase.

APPENDIX I

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Albasing, E. L., "Error in Digital Solution of Linear Problems" Error in Digital Computation, Vol. I (Proc. Advanced Sein. conducted by Math Res. Center, U.S. Army, Univ. Muse., Madison, Wis. 1964) pp. 131-184, Wiley, New York, 1965.
2. Bakusinskii, A. B., "A Method of Solving "Degenerate" and Almost "Degenerate" Linear Algebraic Equations" Z. Vycisl Mat., Mat. Fig. 3 (1963) 1113-1114.
3. Belostochii, A. J., "An Estimate for the Precision of Approximate Solutions of a System of Linear Algebraic Equations" Z. Vycisl Math. i Mat. Fig. 5 (1965) 112-114
4. Belostochii, A. J., "On a Method of Solving Equations Probl. Numer Math. Comp. Tech." (Russian) pp. 71-81 Goserdorstv Nauchno-tehn Izdat Mosinostr Lit., Moscow, 1963.
5. Bergmann, Wilfred, "Ein Algorithm zur Formulierung Von Iterations Verfahren in der Linearen Algebra", Wiss Z. Techn. Univ. Dresden 12: (1963) 114-116.
6. Blanc, C., "Sur L'Estimation des Erreurs D'Arrondi Information Processing" pp. 54-57 UNESCO Paris; R. Oldenbourg Minich Butlerworth, London, 1960
7. Carpenter, J., "Eliminations Ordonnees Un Processus Dimmuant le Volume Des Calculs Dans la Resolution des Systemes Lineaires a Matrice Creuse", Troisieme Cong. de Calcul et de Traitement de l'Information AE (Alti, pp. 63-71 Dinod, Paris, 1964).
8. Chartres, B. A., "Adaptation of the Jacobi Method for a Computer with Magnetic Tape Backing Store" Comput. J5 (1962): 51-60
9. DeMeersman R., and Schotsmans, L., "Note on the Inversion of Symmetric Matrices by the Gauss-Jordon Method: ICC Bull. 3 (1964): 152-155
10. D'Sylva, D. and Miles, G. A., "The SSOR Iteration Scheme for Equations with σ_1 Ordering" Compt. J6 (1961): 43-60
11. Dufour, H. M., "Resolution des Systemes Lineaires Par la Methode des Residus Conjugues" Bull Geodesique (N.S.) No. 71 (1964): 65-87
12. Dufour, H. M., "Resolution des Grands Systemes Lineaires"; Methodes Iterative"; "Methodes par Elimination"; "Essai Divne Methode Synthetisant les deux Points de vue cas de l'Equation de LaPlace" Deux Congr. Assoc. Francaise Calcul et Traitement Information (Paris 1961): pp. 15-37 (Gauthier Villars' Paris, 1962)

13. Dwyer, P. S., "Matrix Inversion with the Square Root Method" Technometrics 6 (1964): 197-213
14. Fiedler, M. "Estimates and Iteration Procedures for Proper Values of Almost Decomposable Matrices (Russian Summary), Czech Math J.14 (87)(16)(64): 593-608
15. Fiedler, M. "Some Remarks on Numerical Solution of Linear Problems" Appl. Math 10 (1965): 190-193
16. Fiedler, M. "Some Applications of the Theory of Graphs in Matrix Theory and Geometry" Theory of Graphs and Its Applications (Proc. Symps. Smolenice, 1963) pp. 37-41. Publ. House Czechoslovak Acad. Sci. Prague 1964
17. Fiedler, M. "On Aggregation in Matrix Theory and Its Application to Numerical Inverting of Large Matrices", Bull. Acad. Polon. Sci. Ser. Sci. Math Astron. Phy. 11 (1963): 757-759
18. Fiedler, M. "On Inverting Partitioned Matrices" Czechoslovak Math. J.13 (88)(1963): 574-586
19. Feldmann, H. "Ein Hinreichendes Konvergenzkriterium und Eine Fehlerabschätzung für die Iteration in Einzelschritten bei Linearen Gleichungssystemen, Z. Angew Math Mech. 41 (1961):515-516
20. Focke, Joachim "Über die Kondition Linearer Gleichungssysteme," Wiss. Z. Karl Marx Univ. Leipzig Math. Nat. Reihe 11 (1962): 41-43
21. Fornheim, L. "Determination of Large Parallel Pipeds", SIAM Rev. 4 (1962): 223-226
22. Frank, Pierce "Sur La Plus Courte Distance d'une Matrice Donnée à l'Ensemble des Matrices Singulières (R. Acad. Sci. Paris 256 (1963): 3799-3801
23. Frank, Pierce, "Sur la Distance Minimale d'une Matrice Régulière Donnée au lieu des Matrices Singulières", Deux Cong. Assoc. Française Calcul et Traitement d'Information (Paris, 1961):pp.55-60 Gauthier, Villars Paris 1962
24. Golub, G. H. and Varga, R. S. "Chebyshev Semi-Iteration Methods Successive Over-Relaxation Iterative Methods, and Second Order Richardson Iterative Methods", I and II Numer Math. 3 (1961): 147-156 157-168
25. Golub, G. H. and Varga, R. S. "Bounds for the Round-off Errors in the Richardson Second Order Method" Nordisk Tidsskrift Informatics Behandling 2 (1962): 212-223

26. Harary, Frank "A Graph Theoretic Approach to Matrix Inversion by Partitioning Numer" Math 4 (1963): 128-135
27. Heinrich, H. "Bemerkung zu Eimen Konditionsmass fur Lineare Gleichungssysteme" Z. Angew Math Mech. 43 (1963): 568
28. Kljuev, V. V. Kokovkin and Scerbak, N. K., "On the Minimization of the Number of Arithmetic Operations for Solving Linear System of Equations" Z. Vycisl Mat. Fig. 5 (1965): 21-33
29. Melbasinshi, A. S., "A Generalization of A. M. Ostrowski Theorum on Iteration Processes", Vestnik, Moskov Univ. Ser. I Math. Mech. 1960 # 5: 40-48
30. Kraustsengl, Rudolf, "A Note on Increasing the Convergence Rate of a Simple Iterative Process for the Solution of Linear Systems" Apl. Mat. 9 (1964): 399-409
31. Liebl, P. "Einige Bermerkunegen Zur Numberischen Stabilitat von Matrigenitestationen" Apl. Math 10 (1965): 249-254
32. Moksmenko, J. F. "Some Topological Methods of Studying Linear Structures" Szv. SSSR Iehm. Kibernet, 1964, # 2: 102-113
33. Mihajlover, B., "A Practical Spectral Methof for Approximate Numerical Solution of a System of Linear Equations" Bull. Soc. Math. Phys. Serber II (1959): 145-150
34. Mirsky, L., "Inequalities and Existence Theorems in the Theory of Matrices" J. Math. Anal. Appl. 9 (1964): 99-118
35. Narsul, A. B., "Improving the Converging of Methods of Successive Approximation for Linear Equations" Dokl. Akad. Mauk SSSR 158 (1964): 279-280
36. Oettle, W. and Prager, W., "Compatibility of Approximate Solution of Linear Equations with Given Error Bounds for Coefficients and Right-Handed Side" Numer. Math 6 (1964): 405-409
37. Osborne, E. E., "On Pre-Conditioning of Matrices", J. Assoc. Compt. Math. 7 (1960): 338-345
38. Petryshyn, W. V., "On the Extrapolated Jacobi or Simultaneous Displacements Method in the Solution of Matrix and Operator Equations", Math. Comp. 19 (1965): 37-55
39. Petryshyn, W. V., "On the Inversion of Matrices and Linear Operations" Proc. Amer. Math. Soc. 16 (1965): 893-901
40. Porter, S., "The Use of Linear Graphs in Gaussian Eliminations" SIAM Rev. 3 (1961): 119-130

41. Rigal, J. "Nombre de Conditions et de Significations en Analyse Matricielle", Deux. Congr. Assoc. Franchise Calcul Traitement Information (Paris, 1961): 47-53, Causthier - Villars, Paris, (1962)
42. Roppert, J. "Ein Schnell Konvergerendes Iterations - Verfahren zur Matrix Inversion", Metrika 8 (1964): 152-154
43. Samanskiv, V. E., "Convergence of Iterative Processes" Ukrain Mat. Z 13 (1961) # 3: 113-115
44. Schmidt, J. W., "Ausgangvektoren fur Monotone Iterationen bei Linearen Gleichungssysteme Numer." Math 6 (1964): 78-88
45. Schmidt, J. W., "Konvergenzbeschleunigung bei Monotonen Vektorfolgen" Act a Math. Acad. Sci. Hungr. 16 (1965): 221-229
46. Schneider, Hans (Editor), Recent Advances in Matric Theory The University of Wisc. Press, Madison, Wisc. 1964 XI: 142
47. Schroder, J., "Computing Error Bounds in Solving Linear Systems" Math. Comp. 16 (1962): 323-337
48. Sosis, P. M. "On the Solutions of Systems of Linear Equations with well Defined Matrix Structures on Electronic Computers" Z Vycisl. Nat. i Nat. Fig. 3 (1963): 777-780
49. Tornheim, Leonard "Convergence of Multipoint Iterative Methods" J. Assoc. Comp. Mach. 11 (1964): 210-220
50. Varga, Richard S. "Iterative Methods for Solving Matrix Equations" Amer. Math. Monthly 72 (1965), #2, Pt. II: 67-74
51. Voevodin, V. V., "The Convergence of the Orthogonal Power Method: Z. Vycisl. Mat. i Mat. Fig. 2 (1962): 529-536
52. Wilkinson, J. H. "Instability of Elimination Method of Reducing a Matrix to Tridiagonal Form" Comp. J. 5 (1962): 61-70
53. Wilkinson, J. H. "Error Analysis of Direct Methods of Matrix Inversion" J. Assoc. Comput. Mach. 8 (1961): 281-330
54. Wynn, P. "Acceleration Techniques for Iterated Vector and Matrix Problems" Math. Comp. 16 (1962): 301-302
55. Zenken, O. V. "Some Remarks on the Stability of Iteration Processes" A. Vycisl. Mat. i Mat. Fig. 4 (1964): 745-748
56. Zurmuhl, R. "Zur Iterativen Behandlung von Matrizeneigenwerten" Wiss Z. Techn. Univ. Dresden 10(1961): 1041-1043